DOCUMENT RESUME

ED 368 183 FL 021 929

AUTHOR Ross, Steven; Hua, Te-Fang
TITLE An Approach to Gain Score Dependability and Validity for Criterion-Referenced Language Tests.
PUB DATE Mar 94
NOTE 28p.; Paper presented at the Annual Language Testing Research Colloquium (16th, 1994).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Achievement Gains; *Criterion Referenced Tests; English (Second Language); Higher Education; Language Proficiency; *Language Tests; Scores; *Test Content; *Test Reliability; *Test Theory

ABSTRACT

A general issue related to language program development involves the empirical rationalization of cut score decisions in criterion-referenced language tests. Cut score dependability focuses on the consistency of the decisions in repeated testing or the assessment of language learner performances. In this case, the issue is to determine the optimal index of gain score dependability in the pre-instruction and post-instruction approach to assessing the language learning gains. This paper examines an approach used in assessing gain score dependability in which the optimal index of gain score dependability is derived from examining the cut score dependability of the pre-instructional administration of the criterion-referenced test as well as the post-instructional criterion-referenced test, in relation to differences in the ratio of pre- and post-instruction variances. The database comes from a pre-instruction administration of a university-level academic listening test followed by a counterbalanced post-instruction administration of an alternate form of the same test after one semester of instruction. Results indicated only moderate gain score dependability, because the cut score was close to the mean proportion correct on the post-instructional test. (MDM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 368 183

AN APPROACH TO GAIN SCORE DEPENDABILITY AND VALIDITY FOR CRITERION-REFERENCED LANGUAGE TESTS

Steven Ross
University of Hawai'i Manoa

Te-Fang Hua
East West Center and University of Hawai'i Manoa

Abstract

Much of the recent work on criterion-referenced language testing addresses the issues of item writing and cut score dependability. Criterion-referenced item writing is centrally concerned with determining the content congruence and learnability of each item's content. Cut score dependability focuses on the consistency of decisions in repeated testing or the assessment of language learner performances. A more general issue related to language program development also involves empirical rationalization of cut score decisions. In this case the issue is of determining the optimal index of gain score dependability in the pre-instruction and post-instruction approach to assessing the language learning gains. The present paper examines a commonly used approach to assessing gain score dependability. The optimal index of gain score dependability is derived from examining the cut score dependability of the pre-instructional administration of the criterion-referenced test as well as the post-instructional criterion-referenced test, in relation to differences in the ratio of pre and post instruction variances. The database for the present paper comes from a pre-instruction administration of an academic listening test followed by a counterbalanced post-instruction administration of an alternate form of the same test after one semester of instruction. The subjects were 213 advanced ESL learners at a large American university English language institute.

FL021929

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Te-Fang Hua
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)"

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

BEST COPY AVAILABLE

CRT in Language Testing

The need for making language tests optimally useful for the assessment of second language instructional programs has been a point of discussion for more than a decade (Cziko 1981; Henning 1982). A trend toward designing tests to assess the effects of instruction has recently gathered momentum in the field of language testing. The advantages of using criterion-referenced tests in language programs stem from their better fit to the content of tasks, objectives, and linguistic structure included in second language syllabus, and their potential to more accurately indicate changes in proficiency as a direct result of instruction. Criterion-referenced tests also potentially provide a more dependable basis for program evaluation (Brown, 1991).

Recent discussions of the advantages of criterion-referenced testing in the field of language testing have tended to dwell on test building procedures (Brown, 1991). A key notion in criterion-referenced test making is the difference index, which is used to assess an individual item's capacity to reflect learners' gain in skill or knowledge (Hudson and Lynch, 1984; Hudson, 1993). The difference index for an individual item is the percentage correct on the pre-instruction administration of the test subtracted from that item's post-instruction percentage correct for the same group of learners.

The advantages of building criterion-referenced tests in program development and assessment are numerous. By designing the content of the test items to be optimally congruent with the

instructional syllabus, the potential for item content validity is maximized (Brown, 1991). Also, if there are demonstrable gains on some items, but not on others, the degree of learnability of subcomponents of the language teaching syllabus can be better examined and revised. In contrast to using standardized or norm-referenced measurements as post-instruction criteria for program evaluation, the criterion-referenced approach offers the advantage of detecting individual differences in change vis a vis the content of the syllabus. Norm-referenced tests, in contrast, tend to cover a much wider range of items by concurrently sampling larger domains of linguistic knowledge (Hudson and Lynch, 1984; Brown, 1991).

The introduction of criterion-referenced testing to the field of language testing has yet to date tended to dwell on the item making and interpretation process. There are implicit assumptions about the dependability and validity of total score gain as the direct result of instruction in the pre-instruction and post-instruction interpretation of difference indices for individual items. The direct comparison of pre- and post-test total score differences can lead to gain scores that may present problems for determining their reliability and validity (Lord, 1963; cf. Rogosa and Willett, 1983).

Current approaches to criterion-referenced test dependability primarily rely on either comparisons of dichotomous judgements of mastery on two independent administrations of a test (Subkoviak, 1980, 1988; Brown, 1990), or rely on squared-error loss agreement

approaches (Berk, 1984), that detect the proximity of a score to a criterion or cut score along a continuum (Brown, 1990). One advantage of a squared-error loss agreement approach resides primarily in the fact that a single test administration is thought to be adequate for determining the dependability of the decision about individual scores (Brennan, 1980, 1984; Brown, 1990).

The squared-error loss agreement approach, usually in the form of phi, calculated at a given cut-score (lambda), provides a dependability index for each of the criterion-referenced test administrations in the pre-instruction and post-instruction scheme appropriate for the assessment of instructional programs. Here lambda is a pre-determined standard for mastering set for both pre-test and post-test. Lambda can be a different proportion on each of the test administrations.

Figure 1 Phi Lambda Index of CRT Dependability

$$\Phi(\lambda) = 1 - \frac{1}{k-1}\left[\frac{\bar{X}_p(1-\bar{X}_p) - S_p^2}{(\bar{X}_p-\lambda)^2+S_p^2}\right]$$

Where:

lambda is the cut score expressed as a proportion

k is the number of items on the test

$\bar{X}_p$ is the mean of proportion scores

S_p is the standard deviation of proportion scores

The cut score dependabilities for the pre-instruction measure and the post-instruction measure do not provide information about

the extent and reliability of pre-to post test gains. For this reason, an elaboration of the criterion-referenced model is warranted - one that can address the dependability and ideally, the validity of instructional gains relative to the cut-scores utilized in the criterion-referenced approach to language testing.

The present study addresses the issue of criterion-referenced gain score interpretation in light of dependability and validity issues. Our focus is on integrating dependability indices with pre-instruction and post-instruction variances on total test scores typically used in academic skill-building instructional programs. Criterion-referenced language test designers could potentially benefit from the experience of test analysts from other areas of educational measurement who have tackled the problem of linking gain scores, or changes in ability before and after instructional programs, to external criteria. The approach used in gain score validity analysis is to link achievement score differences with auxiliary criteria known to assess the same traits as those thought to be developed through instruction (Gupta et al, 1988). The essential difference is that the analysis of gain scores for norm referenced tests have assumptions based on internal consistency and small standard errors of measurement while criterion-referenced tests assume skewedness on pre-and-post instruction distributions, and a pre-set definition of mastery. Brennan, 1984; Hudson and Lynch, 1984; Brown, 1991).

For the practical implementation of criterion-referenced testing in intensive language programs, where there is an explicit

assumption that short term gains will accrue as the direct result of instruction, an implicit assumption is that observed gains are dependable and valid in relation to relevant criteria. The methodology for assessing the effect of instructional programs utilizing criterion-referenced assumptions, however, has not been examined extensively in the language testing literature. The examination of gain scores in terms of mean differences implies a familiar and straight forward approach to assessing instructional effect size. Individual pre-test and post-test scores can be simply collated and a matched t-test can be used to assess the observed mean gain in relation to the null hypothesis. An analogous approach that retains the familiar conception of dependability or reliability, expresses the observed gain after instruction in terms of a magnitude ranging form zero to unity. The sections below explicate how the dependability of gain scores can be used in a criterion-referenced context.

The analysis of gain scores has been conducted in a variety of ways, but one approach that is relevant to criterion-referenced language testing is one that incorporates changes in the distribution of relative variances on pre-instruction and post-instruction measurements for the same cohort of students. Zimmerman and Williams (1982) and Williams, Zimmerman and Mazzagatti (1987) suggest an index of gain score reliability that incorporates the magnitude of the changes from pre-to-post-test in relation to changes in the ratio of pre-and post-test variances. In their approach, the reliability of the gain is greatest when the

pre-test and post-test are internally consistent and show a low correlation. Language testers familiar with gain score reliability will recognize the approach used widely in educational psychology. This approach is based on the internal consistency of the pre-test and post-test instruments, and their correlation.

Figure 2 Internal consistency and correlation-based gain reliability

$$r_{dd} = \frac{r_{xx} + r_{yy} - 2r_{xy}}{2 - 2r_{xy}}$$

Where:

r_{xx} is the internal consistency of the pre-test

r_{yy} is the internal consistency of the post-test

r_{xy} is the product moment correlation between the two tests

Zimmerman and Williams (1982), Rogosa and Willet (1983), and Williams, Zimmerman and Mazzagatti (1987) discuss modifications of the internal consistency and correlation based approach to gain reliability that are optimally sensitive to changes in score distributions from pre-instruction to post-instruction. They add ratios of standard deviation terms (theta, below) to make a product of internal consistency and the ratio of pre-test and post-test score distributions. The Zimmerman and Williams modification (Figure 3) makes explicit the assumption that greater variation among learners is expected before instruction relative to variation after instruction. Figure 3 shows the Zimmerman and Williams (1982)

modification of the internal consistency and correlation-based gain reliability.

Figure 3 Gain score reliability

$$Gain\ Rel = \frac{(\theta_1 r_{xx}) + (\theta_2 r_{yy}) - 2r_{xy}}{\theta_1 + \theta_2 - 2r_{xy}}$$

where:

theta $_1$ is the ratio of pre-to-post-test standard deviations

theta $_2$ is the ratio of post-to-pre-test standard deviations

r_{xx} is the internal consistency of the pre-test

r_{yy} is the internal consistency of the post-test

r_{xy} is the product-moment correlation between the two tests

In order for the gain score reliability concept to apply to the criterion-referenced test dependability interpretations, some adaptations are necessary. By replacing the internal consistency estimates for the pre-instruction administration of the criterion-referenced test with a squared-error loss agreement coefficient *phi*, fixed at a cut score *lambda* for each of the pre- and post-test administrations, the Williams, Zimmerman and Mazzagatti (1987) approach can be adapted to assess gain score dependability. Here, pre-instruction criterion- referenced measures are used as a baseline for language learning gains as indicated on post-instructional criterion-referenced measures for the same

cohort of learners. This approach is premised on there being a cut score on both the pre-instruction and post-instruction versions of the criterion-referenced tests. Figure 3 shows the modification of the norm-referenced approach to gain score reliability to suit the conditions of criterion-referenced gain score dependability.

Figure 4 Gain Score Dependability

$$Gain\ Dependability = \frac{(\theta_1 \Phi(\lambda_x)) + (\theta_2 \Phi(\lambda_y)) - 2r_{xy}}{(\theta_1 + \theta_2) - 2r_{xy}}$$

Where:

theta $_1$ is the ratio of pre-to-post-test standard deviations

theta $_2$ is the ratio of post-to-pre-test standard deviations

phi(lambda $_x$) is the squared-error loss agreement on the pre-test

phi(lambda $_y$) is the squared-error loss agreement on the post-test

r_{xy} is the product-moment correlation between the two tests

Table 1-3 show gain score dependability for criterion-referenced tests. Dependability is calculated for pre-test with post-test correlations at .1, .3, .5, .7 and .9 and differing cut score dependabilities for the pre-test and post-test administrations of the criterion-referenced test.

Table 1 Gain score dependabilities for theta$_1$ =3 and theta$_2$=.5

rxy=.1	rxy=.3	rxy=.5	rxy=.7	rxy=.9	phi L pre/post
0.95	0.94	0.93	0.92	0.90	0.95/0.95
0.89	0.88	0.86	0.83	0.79	0.90/0.90
0.84	0.82	0.79	0.75	0.69	0.85/0.85
0.79	0.76	0.72	0.67	0.59	0.80/0.80
0.73	0.70	0.65	0.58	0.49	0.75/0.75
0.68	0.64	0.58	0.50	0.38	0.70/0.70
0.63	0.58	0.51	0.42	0.28	0.65/0.65
0.58	0.52	0.44	0.33	0.18	0.60/0.60
0.52	0.46	0.37	0.25	0.07	0.55/0.55
0.47	0.40	0.30	0.17	0.00	0.50/0.50
0.42	0.34	0.23	0.08	0.00	0.45/0.45
0.36	0.28	0.16	0.00	0.00	0.40/0.40
0.31	0.22	0.09	0.00	0.00	0.35/0.35
0.26	0.16	0.02	0.00	0.00	0.30/0.30
0.20	0.09	0.00	0.00	0.00	0.25/0.25
0.15	0.03	0.00	0.00	0.00	0.20/0.20
0.10	0.00	0.00	0.00	0.00	0.15/0.15
0.05	0.00	0.00	0.00	0.00	0.10/0.10
0.00	0.00	0.00	0.00	0.00	0.05/0.05
0.13	0.01	0.00	0.00	0.00	0.05/0.95
0.17	0.05	0.00	0.00	0.00	0.10/0.90
0.20	0.09	0.00	0.00	0.00	0.15/0.85
0.24	0.14	0.00	0.00	0.00	0.20/0.80
0.28	0.18	0.05	0.00	0.00	0.25/0.75
0.32	0.22	0.10	0.00	0.00	0.30/0.70
0.36	0.27	0.15	0.00	0.00	0.35/0.65
0.39	0.31	0.20	0.05	0.00	0.40/0.60
0.43	0.35	0.25	0.11	0.00	0.45/0.45
0.47	0.40	0.30	0.17	0.00	0.50/0.50
0.51	0.44	0.35	0.23	0.04	0.55/0.45
0.55	0.48	0.40	0.29	0.12	0.60/0.40
0.58	0.53	0.45	0.35	0.19	0.65/0.35
0.62	0.57	0.50	0.40	0.26	0.70/0.30
0.66	0.61	0.55	0.46	0.34	0.75/0.25
0.70	0.66	0.60	0.52	0.41	0.80/0.20
0.73	0.70	0.65	0.58	0.49	0.85/0.15
0.77	0.74	0.70	0.64	0.56	0.90/0.90
0.81	0.78	0.75	0.70	0.63	0.95/0.05

Table 2 Gain score dependabilities for $theta_1$=2.5 and $theta_2$=1

rxy=.1	rxy=.3	rxy=.5	rxy=.7	rxy=.9	phi L pre/post
0.95	0.94	0.93	0.92	0.90	0.95/0.95
0.89	0.88	0.86	0.83	0.79	0.90/0.90
0.84	0.82	0.79	0.75	0.69	0.85/0.85
0.79	0.76	0.72	0.67	0.59	0.80/0.80
0.73	0.70	0.65	0.58	0.49	0.75/0.75
0.68	0.64	0.58	0.50	0.38	0.70/0.70
0.63	0.58	0.51	0.42	0.28	0.65/0.65
0.58	0.52	0.44	0.33	0.18	0.60/0.60
0.52	0.46	0.37	0.25	0.07	0.55/0.55
0.47	0.40	0.30	0.17	0.00	0.50/0.50
0.42	0.34	0.23	0.08	0.00	0.45/0.45
0.36	0.28	0.16	0.00	0.00	0.40/0.40
0.31	0.22	0.09	0.00	0.00	0.35/0.35
0.26	0.16	0.02	0.00	0.00	0.30/0.30
0.20	0.09	0.00	0.00	0.00	0.25/0.25
0.15	0.03	0.00	0.00	0.00	0.20/0.20
0.10	0.00	0.00	0.00	0.00	0.15/0.15
0.05	0.00	0.00	0.00	0.00	0.10/0.10
0.00	0.00	0.00	0.00	0.00	0.05/0.05
0.27	0.16	0.03	0.00	0.00	0.05/0.95
0.29	0.19	0.06	0.00	0.00	0.10/0.90
0.31	0.22	0.09	0.00	0.00	0.15/0.85
0.33	0.24	0.12	0.00	0.00	0.20/0.80
0.36	0.27	0.15	0.00	0.00	0.25/0.75
0.38	0.29	0.18	0.02	0.00	0.30/0.70
0.40	0.32	0.21	0.06	0.00	0.35/0.65
0.42	0.34	0.24	0.10	0.00	0.40/0.60
0.45	0.37	0.27	0.13	0.00	0.45/0.55
0.47	0.40	0.30	0.17	0.00	0.50/0.50
0.49	0.42	0.33	0.20	0.01	0.55/0.45
0.52	0.45	0.36	0.24	0.06	0.60/0.40
0.54	0.47	0.39	0.27	0.10	0.65/0.35
0.56	0.50	0.42	0.31	0.15	0.70/0.30
0.58	0.53	0.45	0.35	0.19	0.75/0.25
0.61	0.55	0.48	0.38	0.24	0.80/0.20
0.63	0.58	0.51	0.42	0.28	0.85/0.15
0.65	0.60	0.54	0.45	0.32	0.90/0.90
0.67	0.63	0.57	0.49	0.37	0.95/0.05

Table 3 Gain score dependabilities for $theta_1$=2 and $theta_2$=1.5

rxy=.1	rxy=.3	rxy=.5	rxy=.7	rxy=.9	phi L pre/post
0.95	0.94	0.93	0.92	0.90	0.95/0.95
0.89	0.88	0.86	0.83	0.79	0.90/0.90
0.84	0.82	0.79	0.75	0.69	0.85/0.85
0.79	0.76	0.72	0.67	0.59	0.80/0.80
0.73	0.70	0.65	0.58	0.49	0.75/0.75
0.68	0.64	0.58	0.50	0.38	0.70/0.70
0.63	0.58	0.51	0.42	0.28	0.65/0.65
0.58	0.52	0.44	0.33	0.18	0.60/0.60
0.52	0.46	0.37	0.25	0.07	0.55/0.55
0.47	0.40	0.30	0.17	0.00	0.50/0.50
0.42	0.34	0.23	0.08	0.00	0.45/0.45
0.36	0.28	0.16	0.00	0.00	0.40/0.40
0.31	0.22	0.09	0.00	0.00	0.35/0.35
0.26	0.16	0.02	0.00	0.00	0.30/0.30
0.20	0.09	0.00	0.00	0.00	0.25/0.25
0.15	0.03	0.00	0.00	0.00	0.20/0.20
0.10	0.00	0.00	0.00	0.00	0.15/0.15
0.05	0.00	0.00	0.00	0.00	0.10/0.10
0.00	0.00	0.00	0.00	0.00	0.05/0.05
0.40	0.32	0.21	0.06	0.00	0.05/0.95
0.41	0.33	0.22	0.07	0.00	0.10/0.90
0.42	0.34	0.23	0.08	0.00	0.15/0.85
0.42	0.34	0.24	0.10	0.00	0.20/0.80
0.43	0.35	0.25	0.11	0.00	0.25/0.75
0.44	0.36	0.26	0.12	0.00	0.30/0.70
0.45	0.37	0.27	0.13	0.00	0.35/0.65
0.45	0.38	0.28	0.14	0.00	0.40/0.60
0.46	0.39	0.29	0.15	0.00	0.45/0.55
0.47	0.40	0.30	0.17	0.00	0.50/0.50
0.48	0.41	0.31	0.18	0.00	0.55/0.45
0.48	0.41	0.32	0.19	0.00	0.60/0.40
0.49	0.42	0.33	0.20	0.01	0.65/0.35
0.50	0.43	0.34	0.21	0.03	0.70/0.30
0.51	0.44	0.35	0.23	0.04	0.75/0.25
0.52	0.45	0.36	0.24	0.06	0.80/0.20
0.52	0.46	0.37	0.25	0.07	0.85/0.15
0.53	0.47	0.38	0.26	0.09	0.90/0.10
0.54	0.47	0.39	0.27	0.10	0.95/0.05

A Criterion-Referenced Example

Materials

The criterion-referenced test data used in this study came from an advanced academic listening comprehension course at a large American university English language institute. The content of the criterion-referenced test passages and items matched the syllabus specifications for the advanced course. The test content covered listening skills deemed essential for advanced level English as a Second Language students and came from a needs analysis of critical listening tasks. The 21-item test consisted of six major listening tasks: 1) Linking referring pronouns to full noun phrases, 2) Recognizing cohesive devices, 3) Recognizing supporting factual detail, 4) Determining cause and effect, 5) Comprehending vocabulary in context, and 6) Note taking.

Two parallel forms of the advanced listening test were developed for use as pre-tests and post-tests. Each form of the test consisted of thirteen short lecturettes delivered in a narrative style. Each form of the test took approximately twenty-six minutes to complete (excluding instructions). The short criterion-referenced tests (k=21) were designed to assess specific objectives of the course. Their length reflects the assumption that domain-specific tests can sample skill areas as efficiently as longer assessments (Hudson and Lynch, 1984; Brown, 1991).

The pre-test form was administered in the first week of instruction. The strategy for using pre-test measures was predicated on the assumption that students surpassing the pre-test

cut score before instruction got underway could be exempted from the course. The post-test version was the alternate form. The two forms were periodically switched so that each form would be used as pre-test and post-test in different academic terms. The second test administration came at the end of approximately forty hours of classroom instruction in advanced listening skills.

Different criteria for determining passing scores were used on the two administrations of the test. On the pre-test, the cut-score was set at 90% correct of the twenty-one item test. The cut-score for the post-test was set considerably lower at 60% correct.

Subjects

The test data used in this study were gathered on the total pretest and post-test scores of 213 matriculated students enrolled in the advanced listening course. Test records for the 213 subjects were selected from an archive of test results on the condition that both the pre-test and post-test had been completed. No subjects had missing test scores and were enrolled in the advanced academic listening course during of three semesters (Fall 1991, Spring 1992, Fall 1992). All students enrolled in the advanced listening course were either placed directly into the level by a multi-passage academic listening test and dictation used for placement into a two-level academic listening program, or were promoted from an intermediate-level listening course taken in the previous academic term.

Analysis

The assessment of gain was based on the students' total scores on the pre-instruction and the post-instruction tests. The most direct indication of gain is the difference between pre-test and post-test scores. Each pair of scores is independent. That is, no student took the pre-test or post-test more than once. A paired-t test was calculated to examine the significance of the differences between students' performance on the pretest and the post-test. The squared-error loss agreement dependability index was then computed to detect the dependability of the cut score (90% on the pre; 60% on the post) in determining masters from non-masters. Finally, Williams, Zimmerman and Mazzagatti's (1987) approach to gain score reliability was modified by replacing the internal consistency estimates of pre-test and post-test with the squared-error agreement coefficient in order for it to apply to the criterion-referenced test scheme.

Williams, Zimmerman and Mazzagatti (1987) provide three approaches to gain score dependability. All three were examined in the present study using data from the criterion-referenced test administration. The present discussion will be limited to the simple gain score approach outlined by Williams, Zimmerman and Mazzagatti. The simple gain is dependent on the internal consistency of the pre-test scores. Since in this analysis we are dealing with criterion-referenced assumptions, we replace the pre-test and post-test internal consistency estimates with phi(lambda) estimates.

Gain Dependability Results

The modification of the Williams, Zimmerman and Mazzagatti (1987) approach provides the basis for determining the dependability of the gain observed in the advanced listening course. As can be seen in Table 4 there was greater variance on the pre-test relative to the post-test, indicating that the advanced language learners were more homogeneous after instruction. Mean scores on the two tests were also significantly different (paired t= -6.55, p<.005).

Table 4

$\bar{x}_{xp}$=0.570 *Pre-test mean proportion*	$\bar{x}_{yp}$=0.648 *Post-test mean prop*
S_{xp}=0.16 *Pre-test s.d. proportion*	S_{yp}=0.12 *Post-test s.d. prop*
λ_x=0.90 *Pre-test cut score*	λ_y=0.60 *Post-test cut score*
k_x=21 *Pre-test items*	k_y=21 *Post-test items*

Since the cut scores on the two administrations were different, we observed considerable variation in the dependability of the two criterion-referenced tests. On the pre-instructional administration, where the cut score was set at 90%, the median score of the advanced language learners was well below the criterion, resulting in a high pre-test dependability (Figure 5, below) (*phi-x*).

Figure 5 Pre-test Dependability

$$\Phi_x(\lambda_x) = 1 - \frac{1}{k_x - 1}\left[\frac{\bar{X}_{xp}(1-\bar{X}_{xp}) - S_{xp}^2}{(\bar{X}_{xp}-\lambda_x)^2 + S_{xp}^2}\right] = 0.92$$

The post-instruction test, in contrast, resulted in a mean proportion score very close to the 60% threshold for defining a passing score. The post-instruction dependability (*phi-y*) (Figure 6, below) therefore reflects the decreased dependability associated with making decisions about scores so close to the criterion score for mastery of course content.

Figure 6 Post-test Dependability

$$\Phi_y(\lambda_y) = 1 - \frac{1}{k_y - 1}\left[\frac{\bar{X}_{yp}(1-\bar{X}_{yp}) - S_{yp}^2}{(\bar{X}_{yp}-\lambda_y)^2 + S_{yp}^2}\right] = 0.37$$

The gain score dependability index (GDI) (Figure 7) reflects modest dependability of gain on this post-instructional criterion-referenced test. The small decision dependability on the post-test qualifies the interpretation that the gain on the pre-post comparison was uniform among the advanced language learners in the course.

Figure 7 Gain Score Dependability

$$G_{xy} = \frac{(\theta_1 \Phi_x(\lambda_x)) + (\theta_2 \Phi_y(\lambda_y)) - 2r_{xy}}{(\theta_1 + \theta_2) - 2r_{xy}} = 0.63$$

It is also contingent on the correlation (see Table 5) between the pre-instruction and post-instruction forms of the criterion-referenced tests. In circumstances where we would expect learners to have no knowledge of the trait, the pre-post correlation should be near zero. Gain score dependability is largest when the component criterion-referenced measures are dependable and are based on appropriate cut scores, and when the pre-post correlation is near zero.

Table 5

$\theta_1 = \frac{S_{pre}}{S_{post}} = 1.28$ $\theta_2 = \frac{S_{post}}{S_{pre}} = 0.78$ *Ratio of CRT variances*

$r_{xy} = 0.23$ *Pre-Post Correlation*

For the present advanced listening course data set, we find only moderate gain score dependability, because the cut score is close to the mean proportion correct on the post-instructional test.

Gain Score Validity

The linking of internally valid instructional effects in program development is not particularly new in language testing. Henning (1982; 1988) devised approaches to assessing growth-referenced evaluation for language programs that depend primarily on internally-based definitions of validity. The approach followed here anchors external criteria to the pre-to-post gains. Gain

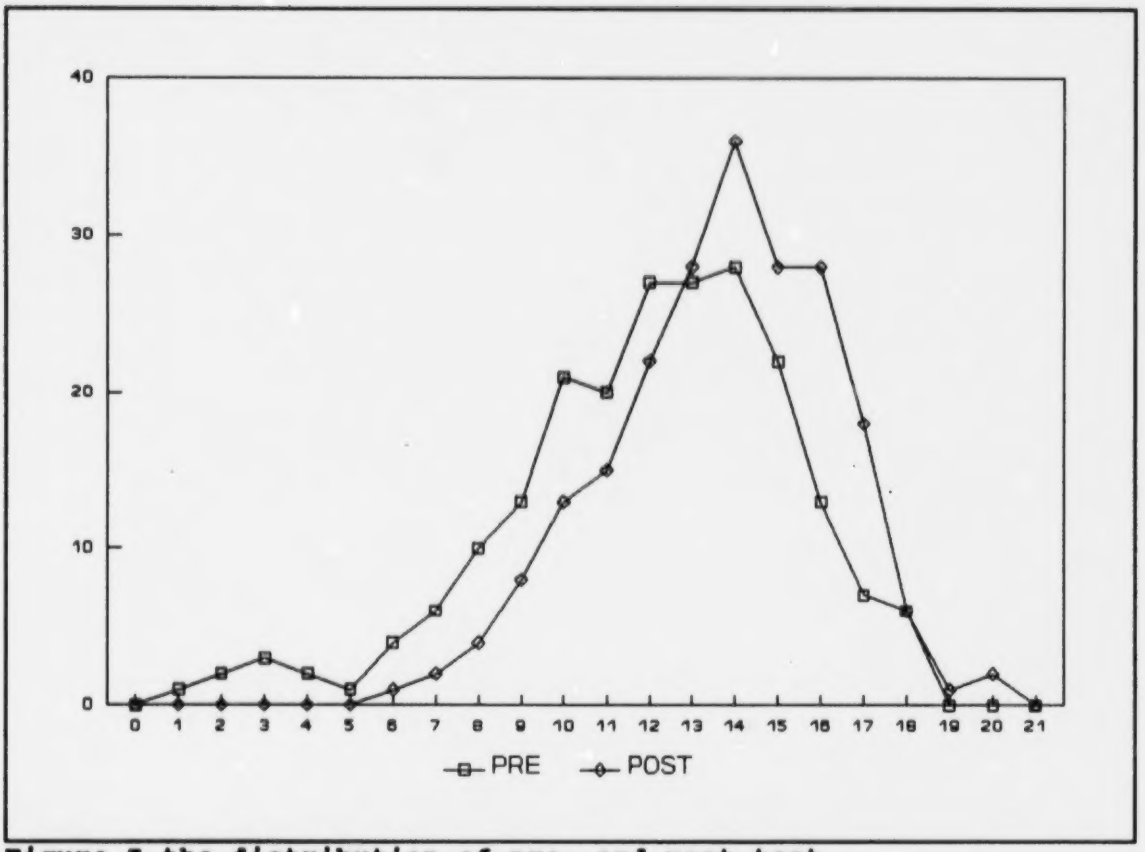

Figure 7 the distribution of pre- and post-test

score validity is premised on the logic that observed gains are relatable to some other external criterion - provided that the external criterion measures the same latent trait as that reflected in the gain scores. Validity is largest when the pre-test and external criterion correlation is zero and the post-test with criterion correlation is very high. We would presume, for example, that before instruction learners' performance would show no correlation with criterion measures. Learner performance after instructional goals have been dependably achieved will reflect the skill or knowledge that can be correlated with an external

criterion assessing the same domain of knowledge or skill. Before instruction, we do not assume that there is any basis for such a correlation. Gupta, Srivastava and Sharma (1988; 1989) define gain score validity as one based on relative magnitudes of pre-post-external criterion correlations.

Figure 8 Gain Score Validity

$$Gv=\frac{(r_{yz}-r_{xz})\theta_1}{\sqrt{\theta_1^2-2r_{xy}\theta_1+1}}$$

where:

r_{xy} is the product-moment correlation between the post-test and the external criterion.

r_{xz} is the product-moment correlation between the pre-test and the external criterion.

r_{yz} is the product-moment correlation between the pre-test and the post-test

The approach adopted here includes the gain score dependability index so as to make the validity index optimally conditioned on the dependability of the pre-to-post instructional gains. We therefore modified the Gupta et al gain score validity approach to suit the conditions of the criterion-referenced approach in the present study, although no external criterion was available for the validation of the observed gains. The exposition below is therefore meant to demonstrate how a gain score validity component can be extended from the gain score dependability approach thus far discussed.

Figure 9 Extended Gain Score Validity

$$GV = \frac{(r_{ys} - r_{xs})\, GDI}{\sqrt{r_{xy} + GDI}}$$

where:

r_{xy} is the pre-instruction with post-instruction CRT correlation

r_{xs} is the pre-instruction with external criterion correlation

r_{ys} is the post-instruction with external criterion correlation

GDI is the gain score dependability index

Ideally, the external criterion would be a parallel form of the criterion-referenced test in a narrowly specified domain related to instructional objectives. The external criterion could itself be validated through conventional means such as multi-trait multimethod approaches used in language testing (Stevenson, 1980; Bachman and Palmer, 1982; Henning and Dandonoli, 1991).

Implications for criterion-referenced language testing

With increasing use of criterion-referenced tests in the evaluation of language teaching programs, there is a concurrent need for determining the extent to which gains can be reliably and validly related to important criteria. The match between the content of instructional programs and the observable gains is ideally related to criteria grounded in the needs of language learners in institutions. The role of external criteria in the validation of the instruction gains is therefore essential for the development of a dynamic language teaching program, especially

in an academic context. The use of external criteria for gain score validity assessment is based on a number of constraints.

In order for there to be optimal assessment of gain in a language teaching program, there are strong assumptions about the degree of variance overlap between pre-instruction and post-instruction measures. In order for gain to be assessed more clearly, the correlation between the pre-to-post instruction measures should be approximately zero. This is a severe assumption for most language teaching programs because academic second language learners initially matriculate with a high degree of proficiency. Since language skills tend to be robustly intercorrelated, the potential for finding specific linguistic subskills that are readily identifiable and teachable also presents a constraint on ascertaining gain score validity.

The gain score validity index is dependent on the external criterion, the pre-to-post instruction correlations and the gain score dependability index. The gain score dependability index is itself subject to the pre-determined cut scores. The basis of the cut scores in criterion-referenced testing (Messick, 1988) is notoriously difficult to justify in absolute terms. Knowing "how much is enough" is typically beyond the standard setters to agree on. In the present study, the 60% on the post-instruction test reflects the intention to make a commonly used threshold in academic settings the minimum criterion for passing. The pre-test cut score of 90% is perhaps less justified. Its function is mainly to allow learners who have been misplaced to demonstrate clear

mastery of the course content, and therefore make themselves candidates for exemption from the required course of instruction. In order to make criterion-referenced tests optimally dependable and equitable to language learners, the pre-instruction cut score should be based on a realistic criterion. One such candidate could be determined by the average proportion of answers correctly answered by several previous cohorts of instructed learners. The pre-instruction cut score could be rationally defined on the average performance of persons who have previously mastered the course content.

Making Criterion-Referenced Tests Work

The crucial element in making criterion-referenced language tests work is the interface of the instructional syllabus and the domain of language knowledge to be taught and assessed. The ideal criterion-referenced test, one that leads to clearly observable gains, is one that samples a knowledge domain that learners do not already possess. This ideal is reflected in the assumptions underlying the gain score validity index - one of which is that pre-instruction and post-instruction correlations are near zero. This assumption, however, will no doubt be extremely difficult to satisfy in instruction programs that focus exclusively structural language teaching and testing - on curricula designed to cover only linguistic knowledge. Assuming that matriculated university English as a second language students already possess advanced knowledge of the language, developing criterion-referenced tests is especially

difficult, and presents a potential criterion-referenced dilemma for language testers. The more discrete point the teaching syllabus and the test content become, the more observable the gains will most likely be. Whether the gains accruing from a narrowly defined domain come at the expense of content validity, and the development of crucial academic skills can only be determined by the continual analysis of the interface of criterion-referenced test content, learner needs, and external criteria.

An alternative to the systemic language teaching and testing approach now prominent in academic language programs is one that integrates procedural knowledge of language-dependent academic research skills into both the teaching syllabus and criterion-referenced testing scheme. An example of such and approach would integrate the teaching of advanced research techniques into a task-based approach to criterion-referenced test design. In this scheme, the content of criterion-referenced tests is not exclusively focused on language structure, reading skills, vocabulary expansion so much as it extends advanced learners' knowledge of specific research procedures in a modern research library. The content of these criterion-referenced tests includes procedural knowledge of advanced reading tasks, e.g., tasks simulating compact disk ROM database search procedures. In contrast to purely systemic language pre-instruction tests, which typically show a high mean and little variance, the integration of procedural research tasks with systemic knowledge tests better fits the low pre-instruction mean score assumptions of currently conceived

criterion-referenced tests. The task-based approach to criterion-referenced testing thus shows potential for accomplishing two important goals. One, to show that the effects of intensive instruction in academic preparation programs can result in tangible gains, and two, that the content of criterion-reference tests can be integrated in academic task simulations that serve to provide comprehensive review of systemic language knowledge while at the same time provide practice in crucial academic research skills for advanced learners.

Acknowledgements

We gratefully acknowledge James Dean Brown, Thom Hudson, and Shuquiang Zhang for their valuable comments on an earlier version of this paper.

References

Bach, L. and Palmer, A. (1982). The construct validation of the FSI Oral Interview. *Language Testing*, 31, 1, 67-86.

Berk, R.A. (1984). Selecting the index of reliability. In Berk, R.A., (Ed.), *A guide to criterion-referenced test construction*. Baltimore: Johns Hopkins University Press.

Brennan, R.L. (1984). Estimating the dependability of the scores. In Berk, R.A., (Ed.), *A guide to criterion-referenced test construction*. Baltimore: Johns Hopkins University.

Brown, J.D. (1990). Short-cut estimators of criterion-referenced test consistency. *Language Testing*, 7, 1, 77-97.

Brown, J.D. (1991). Testing in language programs. Unpublished manuscript.

Cziko, G. (1981). Psychometric and edumetric approaches to language testing: implications and applications. *Applied Linguistics*, 2, 1, 27-44.

Gupta, J.K., Srivastava, A.B.L., and Sharma, K.K. (1988). On the optimum predictive potential of change measure. Journal of Experimental Education 56, 124-128.

Gupta, J.K., Srivastava, A.B.L., and Sharma, K.K. (1989). Estimation of true change using additional information provided by an auxiliary variable. Journal of Experimental Education 57, 143-150.

Henning, G. (1982). Growth-referenced evaluation of foreign language instructional programs. *TESOL Quarterly 16*, 4. 27-44

Henning, G. (1987). A Guide to Language Testing: Development, Evaluation, Research. Rowley, MA: Newbury House.

Henning, G. and Dandonoli ?(1990)

Hudson, T. and Lynch, B. (1984) A criterion-referenced measurement approach to ESL achievement testing. Language Testing 1,2, 171-210.

Hudson, T. (1993) Surrogate indices for item information functions in criterion-referenced language testing. Language Testing 10, 171-191.

Messick, S. (1988) The once and future issue for validity: Assessing the meaning and consequences of measurement. In H. Warner and P. Braum (Eds.) Test Validity. Hillsdale NJ: Lawrence Erlbaum and Associates.

Lord, F. M. (1963) Elementary models for measuring change. In C.W.Harris (Ed.) Problems in Measuring Change. Madison: University of Wisconsin Press.

Rogosa, K.L and Willet, J.B. (1983) Demonstrating the reliability of the difference score in the measurement of change. Journal of Educational Measurement 20, 335-343.

Stevenson, D. (1981) Beyond faith and face validity: The multitrait multimethod matrix and the convergent and discriminant validity of oral proficiency tests. In A. Palmer, P.J.M. Groot and G. Trosper (Eds.) The Construct Validation of Tests of Communicative Competence. Washington DC: TESOL.

Subkoviak, M.J. (1980) Decision-consistency approaches. In Berk, R.A. (Ed.) Criterion-Referenced Measurement:The State of the Art. Baltimore: Johns Hopkins University Press.

Subkoviak, M.J. (1988) A practitioner's guide to computation and interpretation of reliability indices for mastery tests. Journal of Educational Measurement 25, 47-55.

Willliams, R.H., Zimmerman, D.W., and Mazzagatti, R.D. (1987) Large sample estimates of the reliability of simple, residiualized, and base-free gains scores. Journal of Experimental Education 56, 116-118.

Zimmerman, D.W. and Williams, R.H. (1982) Gain scores in research can be highly reliable. Journal of Educational Measurement 19, 149-154.